Article

Classification and Visualisation of Normal and Abnormal Radiographs; a comparison between Eleven Convolutional Neural Network Architectures

Ananda Ananda [1], Kwun Ho Ngan [1], Cefa Karabağ [1], Aram Ter-Sarkisov [2], Eduardo Alonso [2] and Constantino Carlos Reyes-Aldasoro [1]

[1] giCentre, Department of Computer Science, School of Mathematics, Computer Science and Engineering, City, University of London, London EC1V 0HB, UK; Ananda.Ananda@city.ac.uk; reyes@city.ac.uk

[2] CitAI Research Centre, Department of Computer Science, School of Mathematics, Computer Science and Engineering, City, University of London, London EC1V 0HB, UK

Version July 20, 2021 submitted to Sensors

Abstract: This paper investigates the classification of radiographic images with eleven convolutional neural network (CNN) architectures (*GoogleNet, VGG-19, AlexNet, SqueezeNet, ResNet-18, Inception-v3, ResNet-50, VGG-16, ResNet-101, DenseNet-201 and Inception-ResNet-v2*). The CNNs were used to classify a series of wrist radiographs from the Stanford Musculoskeletal Radiographs (MURA) dataset into two classes - normal and abnormal. The architectures were compared for different hyper-parameters against accuracy and Cohen's kappa coefficient. The best two results were then explored with data augmentation. Without the use of augmentation, the best results were provided by Inception-Resnet-v2 (Mean accuracy = **0.723**, Mean kappa = **0.506**). These were significantly improved with augmentation to Inception-Resnet-v2 (Mean accuracy = **0.857**, Mean kappa = **0.703**). Finally, Class Activation Mapping was applied to interpret activation of the network against the location of an anomaly in the radiographs.

Keywords: Wrist Fractures; Radiographic Images; Classification; Convolutional Neural Networks; Class Activation Mapping

1. Introduction

Fractures of the wrist and forearm are common injuries, especially among older and frail persons who may slip and extend the arm to protect themselves [1]. In some cases, the person involved may think that they have not injured themselves seriously and the fractures are ignored and left untreated [2]. These fractures can provoke impairment in the wrist movement [3]. In more serious cases, fractures can lead to complications such as ruptured tendons or long-lasting stiffness of the fingers [4] and can impact the quality of life [5].

Treatment of fractures through immobilisation and casting is an old, tried-and-tested technique. There are Egyptian records describing the re-positioning of bones, fixing with wood and covering with linen [6] and there are also records of fracture treatment in Iron Age and Roman Britain where "skilled practitioners" treated fractures and even "minimised the patient's risk of impairment" [7]. The process of immobilisation is now routinely performed in the Accidents and Emergency (A&E) departments of hospitals under local anaesthesia and is known as *Manipulation under Anaesthesia* (MUA) [8], or closed reduction and casting. MUA interventions in many cases represent a significant proportion of the Emergency Department workload. In many hospitals, patients are initially treated with a temporary plaster cast, then return afterwards for the manipulation as a planned procedure. MUA, although simple, is not entirely free of risks. Some of the problems include bruising, tears of the skin, complications related to the local anaesthetic and there is discomfort for the patients. It should be noted that a large proportion of MUA procedures fail. It has been reported that 41% of Colles' fractures treated with MUA required alternative treatment [9]. The alternative to MUA is open surgery, which is also known as *Open Reduction and Internal Fixation* (ORIF) [10], and can be performed with local or general anaesthesia [11,12] to manipulate the fractured bones and fixate them with metallic pins, plates or screws. The surgical procedure is more complicated and expensive than MUA. In some cases, it can also lead to serious complications especially with metallic elements that can interfere with the tendons and cut through subchondral bones [13,14]. ORIF it is more reliable as a long term treatment.

Despite the considerable research in the area ([8,10,13,15–18]), there is no certainty into which procedure to follow for wrist fractures [19–21]. The main tool to examine wrist fractures is through diagnostic imaging, e.g., X-ray or Computed Tomography (CT). The images produced are observed by highly skilled radiologist and radiographers in search for anomalies, and based on experience, they then determine the most appropriate procedure for each case. The volume of diagnostic images has increased significantly [22], and work overload is further exacerbated by a shortage of qualified radiologists and radiographers as exposed by The Royal College of Radiologists [23]. Thus, the possibility of providing computational tools to assess radiographs of wrist fractures is attractive. Traditional analysis of wrist fractures has focused on geometric measurements that are extracted either manually [24–27] or through what is now considered traditional image processing [28]. The geometric measurements that have been of interest are, amongst others: radial shortening [29], radial length [25], volar and dorsal displacements [30], palmar tilt and radial inclination [31], ulnar variance [24], articular stepoff [26], and metaphyseal collapse ratio [27]. Non-geometric measurements such as bone density [32,33] as well as other osteoporosis-related measurements e.g., cortical thickness, internal diameter, cortical area [34] have also been considered to evaluate bone fragility.

However, in recent years, computational advances have been revolutionised by the use of machine learning and artificial intelligence (AI), especially with *deep learning architectures* [35]. Deep learning is a part of the machine learning methods where input data is provided to a model to discover or learn the representations that are required to perform a classification [36]. These models have a large number of[R3] levels, far more than the input/hidden/output layers of the early configurations and thus considered *deep*. At each level, non linear modules transform the representation of the data from the input data into a more abstract representation [37].

Deep learning has had significant impact in many areas of image processing and computer vision, for instance, it provides outstanding results in difficult tasks like the classification of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [38] and it has been reported that deep learning architectures have in some cases outperformed expert dermatologists in classification of skin cancer [39]. Deep learning has been widely applied for segmentation and classification [40–48].

Deep learning applied system versus radiologists' interpretation on detection and localisation of distal radius fractures has been reported by [49]. Diagnostic improvements have been studied by [50] where deep learning supports the medical specialist to a better outcome to the patient care. Automated fracture detection and localisation for wrist radiographs are also feasible for further investigation [51].

Notwithstanding their merits, deep learning architectures have several well-known limitations: significant computational power is required together with large amounts of training data. There is a large number of architectures, and each of them will require a large number of parameters to be fine tuned. Many publications will use one or two of these architectures and compare against a baseline, like human observers or a traditional image processing methodology. However, a novice user may struggle to select one particular architecture, which in turn may not necessarily be the most adequate for a certain purpose. In addition, one recurrent criticism is their *black box* nature [52–55], which implies that it is not always easy or simple to understand how the networks perform in the way they do. One method to address this *opacity* is through explainable techniques, such as activation maps [56,57] as a tool to visually[R3] explain the localisation of class-specific image regions.

In this work, the classification of radiographs into 2 classes, normal and abnormal, with eleven convolutional neural network (CNN) architectures was investigated. The architectures compared were the following: (*GoogleNet, VGG-19, AlexNet, SqueezeNet, ResNet-18, Inception-v3, ResNet-50, VGG-16, ResNet-101, DenseNet-201 and Inception-ResNet-v2*). This paper extends a preliminary version of this work [58]. Here, we extended the work by applying data augmentation to the two models that provided the best results, that is, ResNet-50 and Inception-ResNet-v2. Furthermore, class activation maps were generated and[R3] analysed.

The dataset used to compare the architectures was the *Stanford MURA (musculoskeletal radiographs)* radiographs [59]. This is a database that contains a large number of radiographs; 40,561 images from 14,863 studies, where each study is manually labelled by radiologists as either normal/abnormal. The radiographs cover seven anatomical regions, namely Elbow, Finger, Forearm, Hand, Humerus, Shoulder and Wrist. This paper focused mainly on the wrist images. The main contributions of this work are the following: (1) an objective comparison of the classification results of 11 architectures, this can help the selection of a particular architecture in future studies, (2) the comparison of the classification with and without data augmentation, which resulted in significantly better results, (3) the[R3] use of Class Activation Mapping to analyse the regions of interest of the radiographs.

The rest of the manuscript is organised as follows. Section 2 describes the materials, that is, the data base of radiographs, and the methods that describe the Deep Learning models that were compared and the Class Activation Mapping (CAM) to visualise the activated regions. The performance metrics of accuracy and Cohen's kappa coefficient are described at the end of this section. Section 3 present the results of all the experiments and the effect of the different hyper-parameters. Predicted abnormality in the radiographic images will also be visualised by using class activation mapping. The manuscript finishes with a discussion of the results in section 4.

2. Materials and Methods

2.1. Materials

The data used to compare the 11 CNNs was obtained from the public dataset MUsculoskeletal RAdiographs (MURA) from a competition organised by researchers from Stanford University [59]. The dataset has been manually labelled by board-certified radiologists between 2001 and 2012. The studies ($n = 14,656$) are divided into training ($n = 13,457$), and validation ($n = 1,199$). Furthermore, the studies have been allocated in groups called abnormal (i.e., those radiographs that contained fractured bones, foreign bodies such as implants, wires or screws, etc.) ($n = 5,715$) or normal ($n = 8,941$). Representative normal cases are illustrated in Fig. 1 and abnormal cases in Fig. 2. The distribution per anatomical region is shown in Table 1. In this paper, the subset of the **wrists** was selected. The cases of normal and abnormal wrist radiographs is presented in Table 2. Notice that these were subdivided into four studies.

Table 1. Distribution of studies of the Stanford MURA (musculoskeletal radiographs) data set [59] for studies of the upper body.

No.	Study	Train		Validation		Total
		Normal	Abnormal	Normal	Abnormal	
1	Elbow	1094	660	92	66	1912
2	Finger	1280	655	92	83	2110
3	Hand	1497	521	101	66	2185
4	Humerus	321	271	68	67	727
5	Forearm	590	287	69	64	1010
6	Shoulder	1364	1457	99	95	3015
7	**Wrist**	**2134**	**1326**	**140**	**97**	**3697**
	Total	8280	5177	661	538	14656

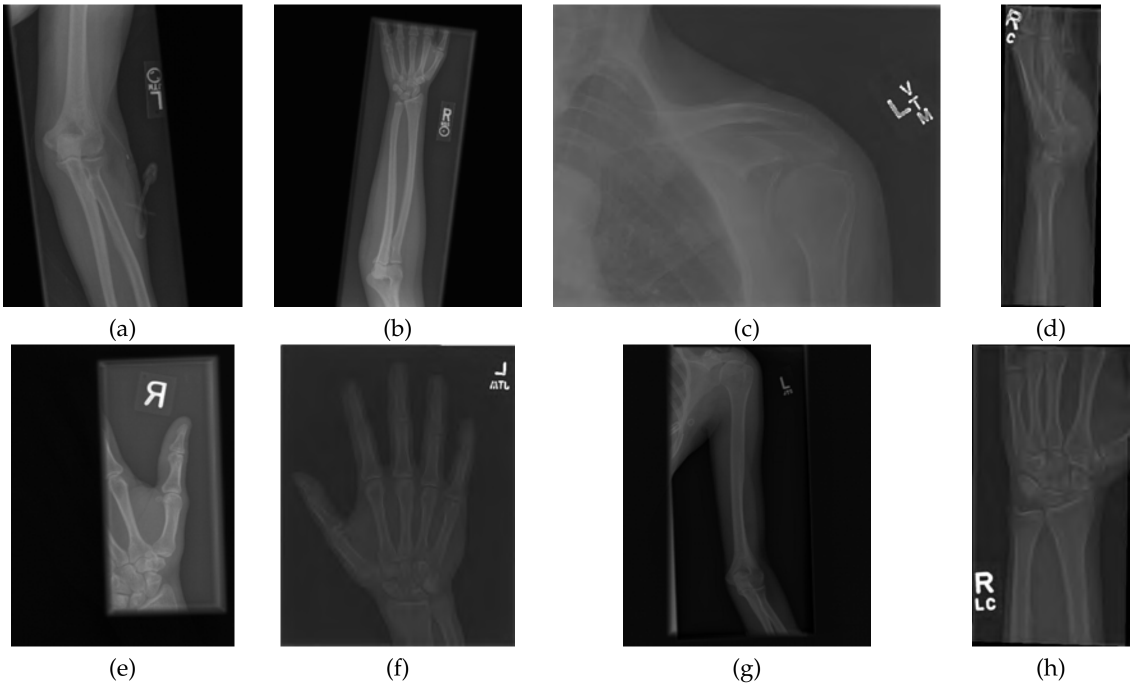

Figure 1. Eight examples of radiographs without abnormalities (considered negative) of the **MU**sculoskeletal **RA**diographs (MURA) dataset [59]. (a) Elbow, (b) Forearm, (c) Shoulder, (d) Wrist (lateral view), (d) Lateral view of Wrist, (e) Finger, (f) Hand, (g) Humerus, (h) Wrist. It should be noted the variability of the images in terms of dimensions, quality, contrast and the large number of labels (i.e.,R for right and L for left), which appear in various locations.

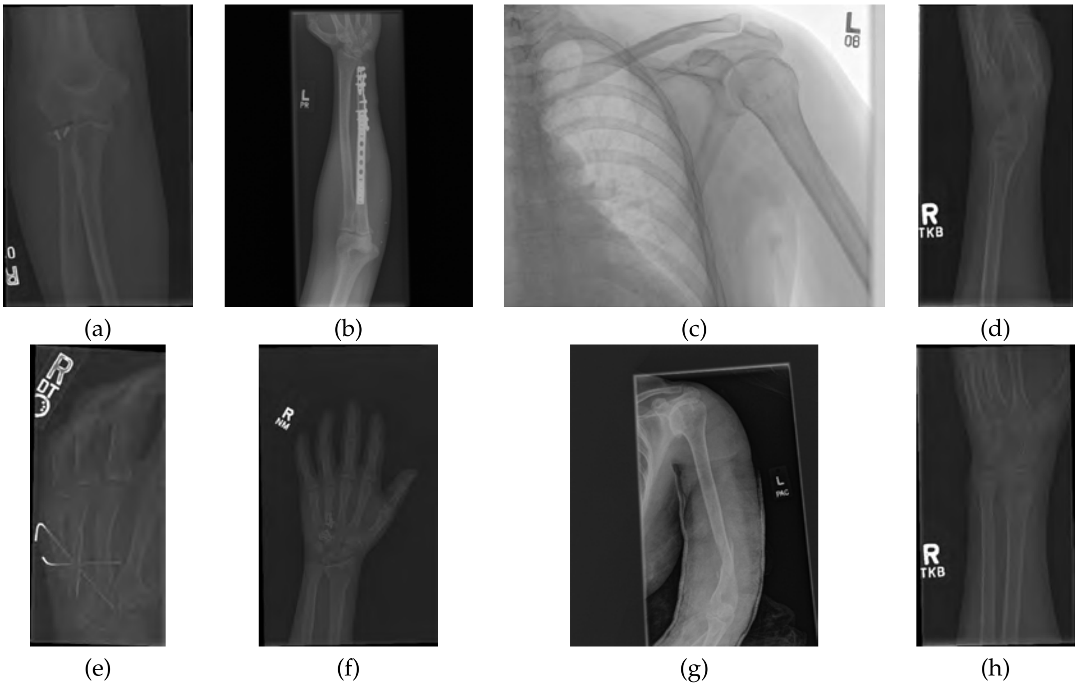

Figure 2. Eight examples of radiographs with abnormalities (considered positive) of the **MU**sculoskeletal **RA**diographs (MURA) dataset [59]. (a) Elbow, (b) Forearm, (c) Shoulder, (d) Wrist (lateral view), (d) Lateral view of Wrist, (e) Finger, (f) Hand, (g) Humerus, (h) Wrist. As for the cases without abnormalities, it should be noted the variability of the images and in addition the abnormalities themselves. There are cases of metallic implants some of which are smaller (a) than others (b), as well as fractures.

Table 2. Details of the number of wrist radiographs. Studies 1,2,3 and 4 refer to a patient visit identifier; each patient may have visited the hospital several times. A positive label, corresponds to abnormal condition, whereas negative corresponds to a normal condition as decided by the expert.

Wrist-Train dataset	**Abnormal**	**Normal**
Study 1	3920	5282
Study 2	64	425
Study 3	3	45
Study 4	0	13
Total	3987	5765
Total Wrist Train Images	9752	
Wrist-Valid dataset	**Abnormal**	**Normal**
Study 1	287	293
Study 2	5	59
Study 3	3	9
Study 4	0	3
Total	295	364
Total Wrist Valid Images	659	
Total Images of Wrist	10411	

2.2. Convolutional Neural Network

Convolutional Neural Networks (CNN) is a type of deep learning [35,36] models. A typical CNN classification model is composed of two key components: first, feature are extracted though a series of convolutional layers with pooling and activation functions. Some modern architectures (e.g. ResNet) will also include batch normalization and/or skip connections to mitigate the problem of vanishing gradient during model training. Next, these features input to one or more fully-connected layers to derive the final classification prediction (e.g. an estimated class probability). These class predictions are used to compute the problem-specific loss.

The input in a CNN, i.e., an image to be classified, can be transformed through the feature extraction layers to form a set of relevant features required by the network. These features can be regarded as the global descriptors of the image. In the fully-connected layers for classification, the relations of the features are learned by an iterative process of weight adjustment. A prediction probability can be deduced at the final layer with the inclusion of an activation function (e.g., softmax function). At the training stage, a loss (e.g. cross entropy loss) is computed between the prediction and the ground truth for weight adjustment during backpropagation. At the evaluation stage, the predicted class can be inferred from most probable class using an argmax function and this can be evaluated against the ground truth for classification accuracy.

A description summary of the applied models used in Table. 3 is as follows: AlexNet [60] is one of the earlier adoptions of deep learning in image classification and has won the ILSVRC 2012 competition by significantly outperformed its next runner up. It consists of 5 layers of convolutions of various sizes and 3 fully connected layers. It also applies a ReLU activation for nonlinearity. GoogleNet (Inception V1) [61] introduced the inception module formed of small size convolutions to reduce trainable parameters for better computational utilisation. Despite a deeper and wider network than AlexNet, the number of parameters for training has reduced from 60 million (Alexnet) to 4 million. VGG [62] is the runner-up in the ILSVRC2014 which was won by GoogleNet in the same year. It utilises only 3x3 convolutions in multiple layers and is deeper than AlexNet. It has a total of 138 million trainable parameters and thus can be computationally intensive during training. ResNet [63] is formed by a deep network of repetitive residual blocks. These blocks are made up of multiple convolution layers coupled with a skip connection to learn the residual based on the previous block. This allows the network to be very deep capable of 100s of network layers. Inception-v3 [64] improves the configuration of the inception module in GoogleNet from a 5x5 convolutional layer in one of the branches to two 3x3 layers reducing the number of parameters. SqueezeNet [65] introduced the fire module which consists of a layer with 1x1 convolution (i.e. squeeze layer) and a second layer with 3x3 convolution (i.e. expand layer). The number of channels into the expand layer is also reduced. This has led to a significant reduction in trainable parameters while maintaining similar accuracy to AlexNet in the ILSVRC 2012 dataset. DenseNet [66] is composed of multiple dense blocks (small convolutional layers, batch normalisation and ReLU Activation). A transition layer with batch normalisation, 1x1 convolution and average pooling is added in between the dense blocks. The blocks are each closely connected with all previous blocks by skip connections. DenseNet has demonstrated a full utilisation of residual mechanism while maintaining model compactness to achieve competitive accuracy. Inception-ResNet-v2 [67] incorporates the advantages of the Inception modules into the residual blocks of a ResNet and achieve even more accurate classification in ILSVRC 2012 dataset than either ResNet 152 or Inception-v3.

Table 3. Details of convolutional neural networks (CNNs) that were used in this work.

No.	Network	Depth	Image Input Size	Reference
1	GoogleNet	22	224-by-224	[61]
2	VGG-19	19	224-by-224	[62]
3	AlexNet	8	227-by-227	[60]
4	SqueezeNet	18	227-by-227	[65]
5	ResNet-18	18	224-by-224	[63]
6	Inception-v3	48	299-by-299	[64]
7	ResNet-50	50	224-by-224	[63]
8	VGG-16	16	224-by-224	[62]
9	ResNet-101	101	224-by-224	[63]
10	DenseNet-201	201	224-by-224	[66]
11	Inception-ResNet-v2	164	299-by-299	[67]

2.3. Experiments

In this work we considered the following eleven CNN architectures to classify wrist radiographs into two categories (Normal / Abnormal): **GoogleNet, VGG-19, AlexNet, SqueezeNet, ResNet-18, Inception-v3, ResNet-50, VGG-16, ResNet-101, DenseNet-201 and Inception-ResNet-v2**. The details of these are presented in Table 3. The training process of the architecture was tested with different numbers of epochs (10, 20, 30), and different mini-batch sizes (16, 32, 64). The experiment pipeline is illustrated in Figure~~.~~[R3] 3. All the architectures were compared under the same conditions, without pre- or post-processing initially except resizing of the initial images to the input size for each architecture as the X-ray images presented different sizes. For instance, the images were resized to 224 x 224 for **ResNet-50** and 299 x 299 for **Inception-ResNet-v2** . In the cases where the input was a 3-channel image, i.e., an RGB colour image, and the input image was in grayscale, this channel was replicated. The dataset was split into 90% for training and 10% for testing. The same hyper-parameters were applied as described in Table 4 and continued in Table 5.

Then, for the two architectures which provided the highest accuracy and Cohen's kappa coefficient (ResNet-50 and InceptionResnet-v2) several modifications were applied regarding, specifically, the use of data augmentation and CNN's training options. The classification with and without augmentation was done to assess the impact that augmentation can have in the results. In addition, visualisation of the network activations with Class Activation Mapping was explored.

2.4. Further processing with data augmentation

For the two best performing architectures, the effect of data augmentation was also be evaluated. The following augmentations have been performed to each of the training images: (1) rotations of (-5 to 5°), (2) vertical and horizontal reflections, (3) shear deformations of (-0.05 to 0.05°) in horizontal

Table 4. Summary of convolutional neural networks (CNNs) hyper-parameters for this work.

1	GoogleNet	Optimiser	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	64	64	64
		Init. Learn. R.	0.01	0.001	0.001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
2	VGG-19	Optimiser	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	64	64	64
		Init. Learn. R.	0.001	0.001	0.001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
3	AlexNet	Optimiser	SGDM	ADAM	RMSprop
		Epoch	50	50	50
		Mini batch size	128	128	128
		Init. Learn. R.	0.001	0.001	0.001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
4	SqueezeNet	Optimiser	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	64	64	64
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
5	ResNet-18	Optimiser	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	64	64	64
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
6	Inception-v3	Optimiser	SGDM	ADAM	RMSprop
		Epoch	10	10	10
		Mini batch size	64	64	64
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
7	ResNet-50	Optimiser	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	64	64	64
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001

Table 5. Summary of convolutional neural networks (CNNs) hyper-parameters for this work (continuation).

8	VGG-16	Optimiser	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	128	128	128
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
9	ResNet-101	Optimiser	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	32	32	32
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
10	DenseNet-201	Optimiser	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	32	32	32
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001
11	Inception-ResNet-v2	Optimiser	SGDM	ADAM	RMSprop
		Epoch	30	30	30
		Mini batch size	32	32	32
		Init. Learn. R.	0.001	0.0001	0.0001
		Momentum	0.9000	-	-
		L2 Reg.	0.0001	0.0001	0.0001

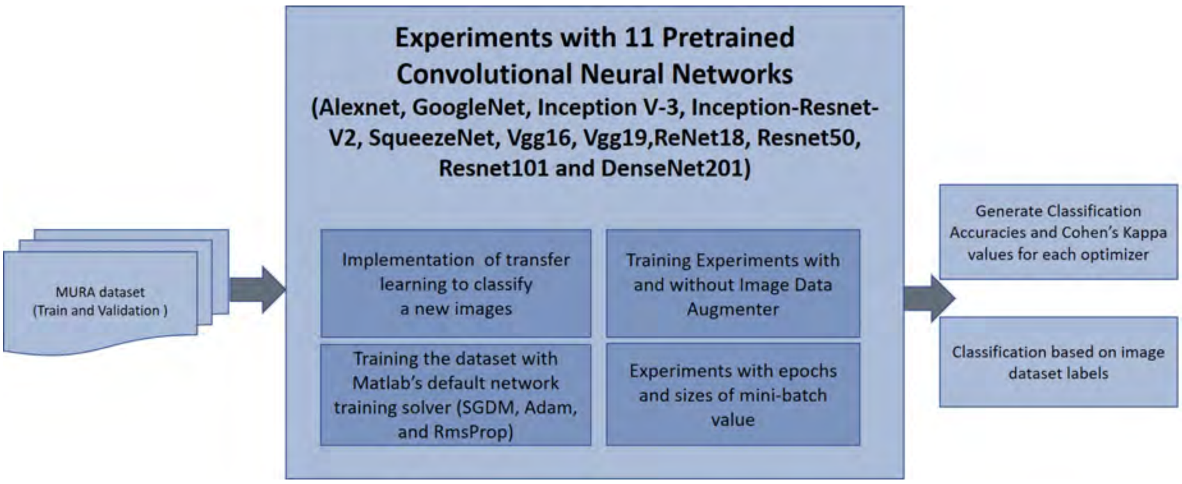

Figure 3. Block diagram which illustrates the classification of the wrist radiographs with 11 different Convolutional Neural Network (CNN) architectures. 9752 images from **MU**sculoskeletal **RA**diographs (MURA) Wrist dataset were used for training CNN architectures and 659 images were used for validation. Two different metrics, Accuracy (Ac) and Cohen's kappa (κ) were computed to assess the performance of 11 pre-trained CNNs. Image data augmentation was used during training and different number of epochs and mini batch sizes were tested.

and vertical directions, and (4) Contrast-limited adaptive histogram equalisation (CLAHE) [68]. Translations were not applied as the training images were captured with a good range of translational shift.

2.5. Class Activation Mapping

Class Activation Mapping (CAM) [56] provides a visualisation of the most significant activation mapping for a targeted class. It provides an indication of what exactly the network is focusing its attention on. Similar to the schematics in Figure 4, the class activation map is generated at the output of the last convolutional layer. In this work, this is represented with a rainbow/jet colour map where the intensity spectrum ranges from blue (lowest activation), green and red (highest activation).

For the two best performing models, the CAM representations were generated at layer "activation_49_relu" for ResNet-50 and "conv_7_bac" for Inception-ResNet-v2 respectively. The CAM maps were up-scaled to the input resolution and overlaid on top of the original radiography for the location of the abnormalities.

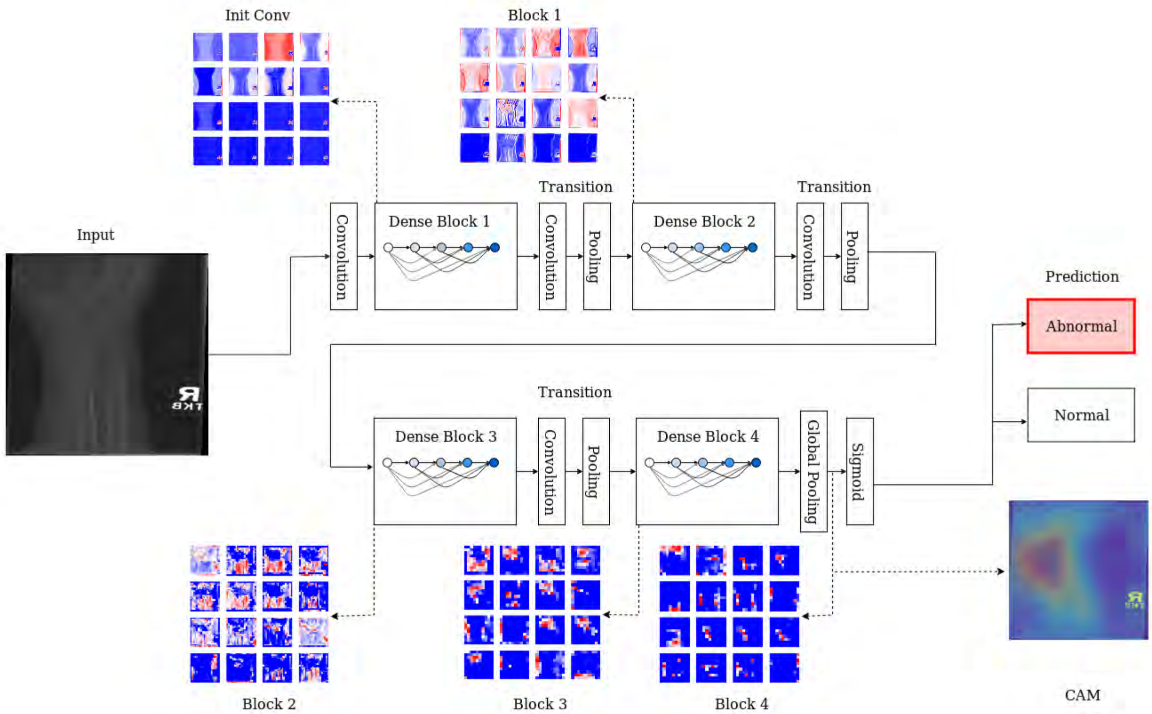

Figure 4. Schematic illustration of the X-ray classification process and class activation mapping through layer-wise activation maps across different dense blocks. At each level, a series of feature maps are generated, the resolution decreases progress through the blocks. Colours indicate the range of activation: blue corresponds to low activation, red for highly activated features. The final output, visualised here using Class Activation Mapping, which highlights the area(s) where abnormalities can be located.

2.6. Performance Metrics

Accuracy (Ac) was calculated as the proportion of correct predictions among the total number of cases examined, that is:

$$Ac = (TP + TN)/(TP + TN + FP + FN), \tag{1}$$

where TP and TN correspond to positive and negative classes correctly predicted and FP and FN correspond to false predictions. Cohen's kappa (κ) was also calculated as it is the metric used to rank the MURA challenge [59,69] and it is considered more robust as it takes into account the possibilities of random agreements. Cohen's kappa κ was calculated in the following way. With

$$Tot = (TP + TN + FP + FN), \tag{2}$$

being the total number of events, the probability of a yes or TP is

$$P_Y = (TP + FP)(TP + FN)/Tot, \tag{3}$$

the probability of a no, or TN is

$$P_N = (FN + TN)(FP + TN)/Tot, \tag{4}$$

and the probability of random agreement $P_R = P_Y + P_N$, then

$$\kappa = (Ac - P_R)/(1 - P_R). \tag{5}$$

2.7. Implementation Details

Experiments were conducted in Matlab R2018b IDE completed with Deep Learning Toolbox, Image Processing Toolbox and Parallel Computing Toolbox. These experiments were conducted using a workstation with a processor from Intel Xeon ® W-2123 CPU 3.60 GHz, 16GB of 2666MHz DDR4 RAM, 500GB SATA 2.5-inch solid-state drive, and NVIDIA Quadro P620 3GB graphic card.

3. Results

The effect of the number of epochs, mini-batch sizing and data augmentation was evaluated on the classification of wrist radiographs in eleven CNN architectures. Table 6 and Table 7 present the aggregated best results for each architecture in prediction accuracy and Cohen's kappa score respectively.

Table 6. Results of accuracy for eleven Convolutional Neural Networks used to classify the wrist images in the MURA dataset. The best results for each row are highlighted in *italics* and the overall the best results are highlighted in **bold**.

No.	CNNs	SGDM	ADAM	Rms Prop	Mean	Epoch	Mini-batch Size
1	GoogleNet	0.650	*0.671*	0.640	0.654	30	64
2	VGG-19	0.680	*0.681*	0.590	0.650	30	64
3	AlexNet	0.674	*0.690*	0.657	0.674	50	128
4	SqueezeNet	0.683	0.657	*0.690*	0.677	30	64
5	ResNet-18	0.704	*0.709*	0.668	0.693	30	64
6	Inception-v3	*0.710*	0.689	0.707	0.702	10	64
7	ResNet-50	0.686	*0.718*	0.716	0.707	30	64
8	VGG-16	0.692	0.713	*0.716*	0.707	30	128
9	ResNet-101	*0.715*	0.706	0.701	0.707	30	32
10	DenseNet-201	*0.733*	0.695	0.722	0.717	30	32
11	Inception- ResNet-v2	0.712	*0.747*	0.710	0.723	30	32
12	ResNet-50 (augmentation)	0.835	0.854	0.847	0.845	30	64
13	Inception-ResNet-v2 (augmentation)	0.842	***0.869***	0.860	*0.857*	30	32

Table 7. Cohen's kappa results from eleven Convolutional Neural Networks used to classify the wrist images in the MURA dataset. The best results for each row are highlighted in *italics* and the overall best results are highlighted in **bold**.

No.	CNNs	SGDM	Adam	Rms Prop	Mean	Epoch	Mini-batch Size
1	GoogleNet	0.373	*0.412*	0.358	0.381	30	64
2	VGG-19	0.433	*0.446*	0.335	0.404	30	64
3	AlexNet	0.420	*0.450*	0.390	0.420	50	128
4	SqueezeNet	0.438	0.390	*0.448*	0.425	30	64
5	ResNet-18	0.474	*0.484*	0.408	0.455	30	64
6	Inception-v3	*0.487*	0.450	0.482	0.473	10	64
7	ResNet-50	0.441	*0.496*	0.494	0.477	30	64
8	VGG-16	0.453	0.491	*0.492*	0.479	30	128
9	ResNet-101	*0.495*	0.475	0.472	0.481	30	32
10	DenseNet-201	*0.524*	0.458	0.507	0.497	30	32
11	Inception-ResNet-v2	0.485	*0.548*	0.484	0.506	30	32
12	ResNet-50 (augmentation)	0.655	*0.696*	0.683	0.678	30	64
13	Inception-ResNet-v2 (augmentation)	0.670	**0.728**	0.711	0.703	30	32

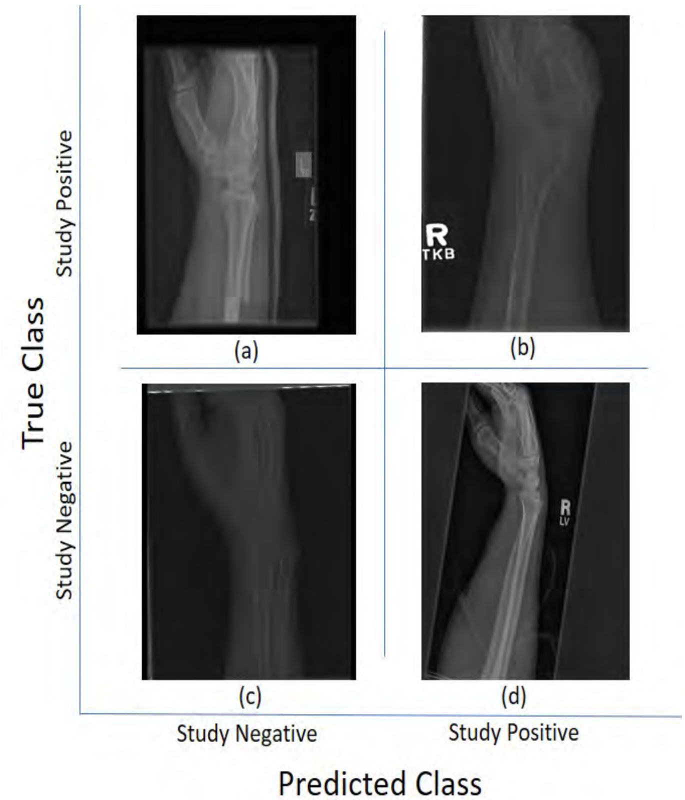

Figure 5. Illustration of classification results for Lateral (LA) views of wrist radiographs. (a) Corresponds to positive (abnormal) diagnosis image but predicted as negative (normal), (b) Abnormal diagnosis and abnormal prediction. (c) Normal diagnosis image and normal prediction. (d) Normal diagnosis and abnormal prediction. Notice that the errors in classification may have been biased by artefact elements on the images.

For the 11 architectures with no data augmentation, Inception-Resnet-v2 performs the best with an accuracy ($Ac = 0.723$) and Cohen's kappa ($\kappa = 0.506$). DenseNet-201 fares slightly lower ($Ac = 0.717$,

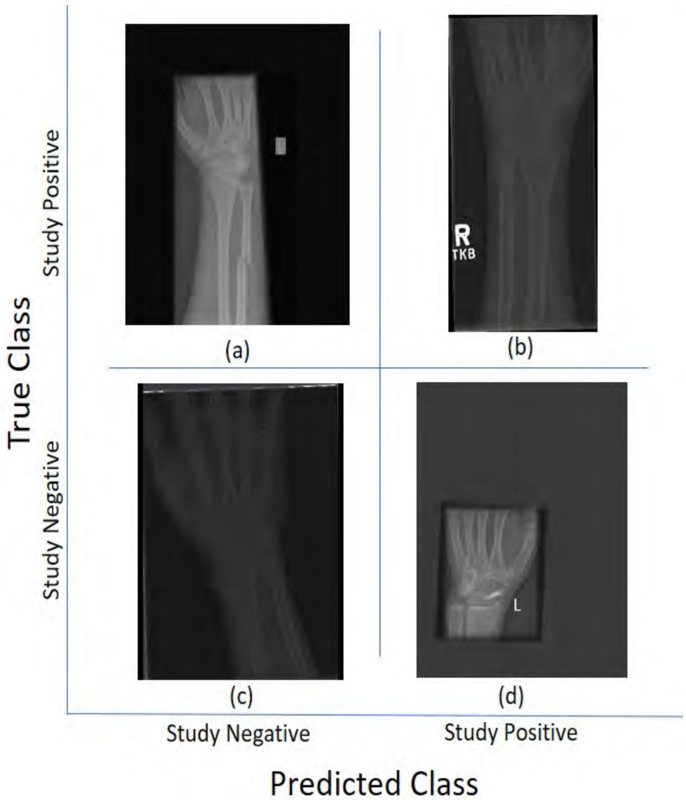

Figure 6. Illustration of classification results for Postero-Anterior (PA) views of wrist radiographs. (a) corresponds to a positive (abnormal) diagnosis image that is predicted as negative (normal); (b) to abnormal diagnosis and abnormal prediction; (c) to normal diagnosis image and normal prediction; and (d) to normal diagnosis and abnormal prediction. Notice again that the errors in classification may have been biased by artefactual elements on the images.

$\kappa = 0.497$). The lowest results were obtained with GoogleNet ($Ac = 0.654$, $\kappa = 0.381$). This potentially indicates better feature extraction with deeper network architectures. Fig. 5 and Fig. 6 illustrate some cases of the classification for Lateral and Postero-anterior views of wrist radiographs.

The comparison between ADAM, SGDM, and RMSprop shows no indicative superiority implying that each of these optimisers were capable of achieving the optimal solution. Incremental change to the number of epochs beyond step 30 yields no improvement in accuracy indicating that the models have converged. The choice of the attempted mini-batches show no difference in results. With data augmentation, the results show significant improvement, e.g., accuracy increases by approximately 19% (up by 0.134) and Cohen's kappa by 39% (up by 0.197) for the Inception-ResNet-v2 architecture.

Class activation maps were obtained and overlaid on top of the representative images in Figures 1 and 2. The CAMs obtained for ResNet50 are shown in Figures 7 and 9 while those for Inception-ResNet-v2 are shown in Figures 8 and 10. In all cases, the CAMs were capable of indicating the region of attention used in the two architectures applied. This is especially valuable for identifying where the abnormalities are in Figure 9 and 10. While both models indicate similar regions of attention, Inception-ResNet-v2 appears to have smaller attention regions (i.e.,more focused) than those in ResNet50. This may indicate a better extraction of features in the Inception-ResNet-v2 leading to better prediction results. Finally, the activation maps corresponding to figures Figures[R1] 5,6 are presented in Figure[R1] 11.

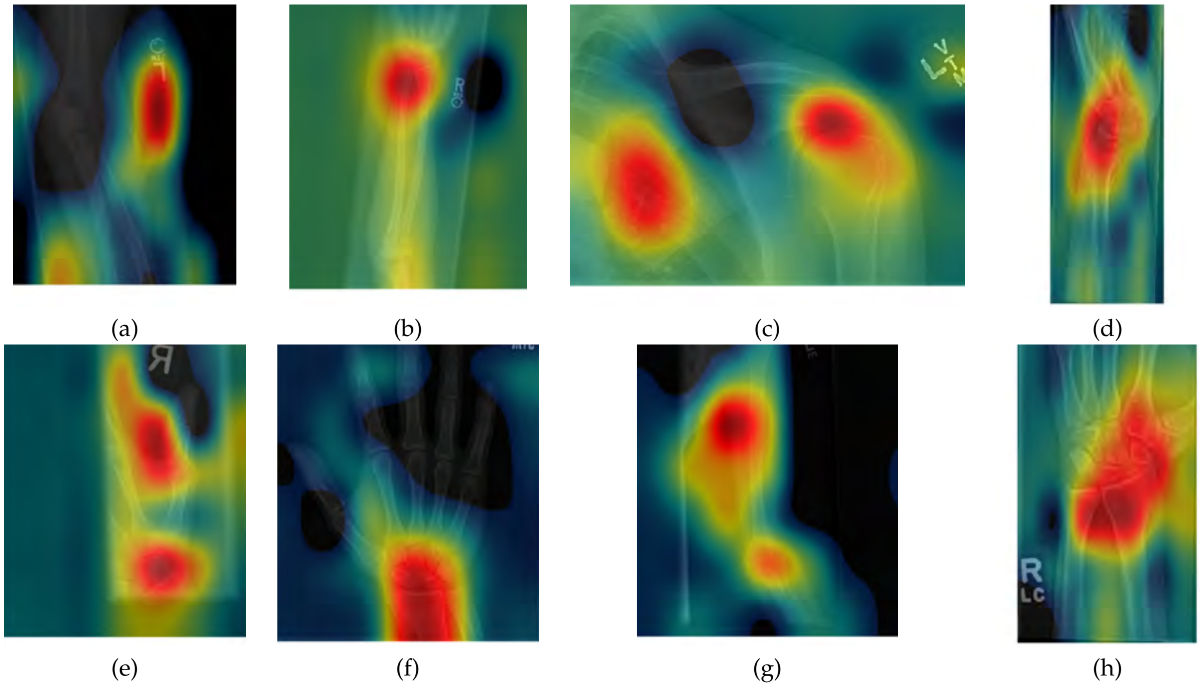

Figure 7. Illustration of Activation Maps overlaid over the eight radiographs without abnormalities of Figure 1 to indicate the regions of the image that activated a **ResNet 50** architecture. (a) Elbow, (b) Forearm, (c) Shoulder, (d) Wrist (lateral view), (d) Lateral view of Wrist, (e) Finger, (f) Hand, (g) Humerus, (h) Wrist. As these cases are positive (no abnormality), the regions of activation are not as critical as those with abnormalities.

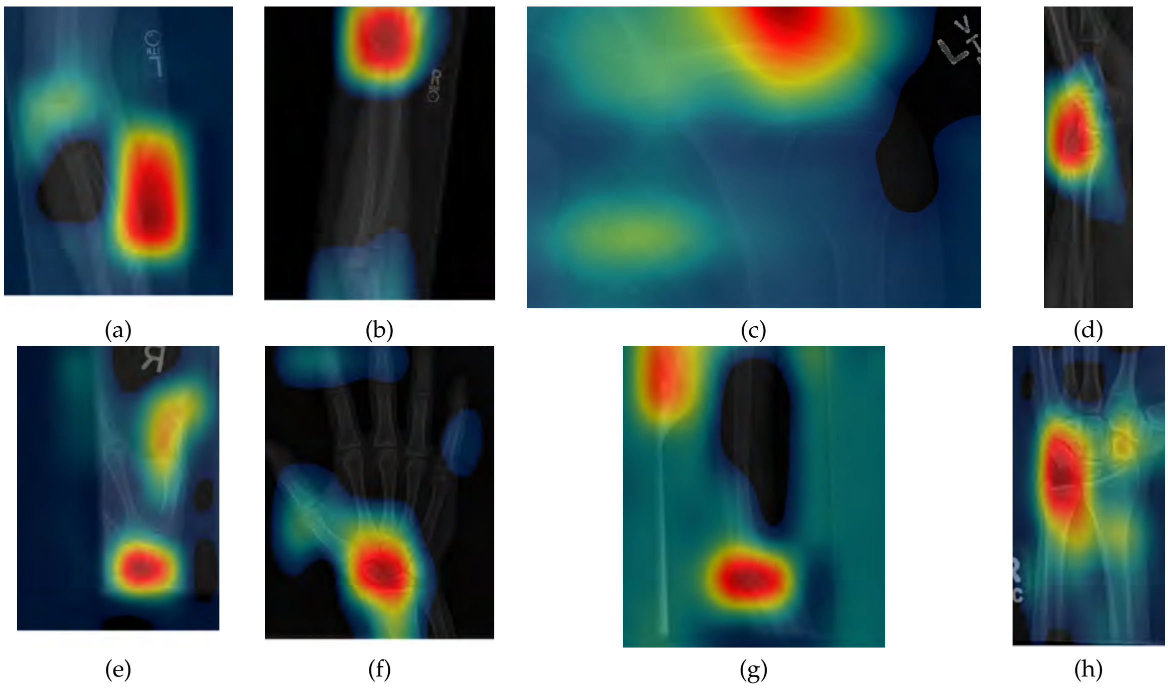

Figure 8. Illustration of Activation Maps overlaid over the eight radiographs without abnormalities of Figure 1 to indicate the regions of the image that activated an **Inception-Resnet-V2** architecture. (a) Elbow, (b) Forearm, (c) Shoulder, (d) Wrist (lateral view), (d) Lateral view of Wrist, (e) Finger, (f) Hand, (g) Humerus, (h) Wrist. It should be noted that the activation regions are more localised than those of the ResNet 50.

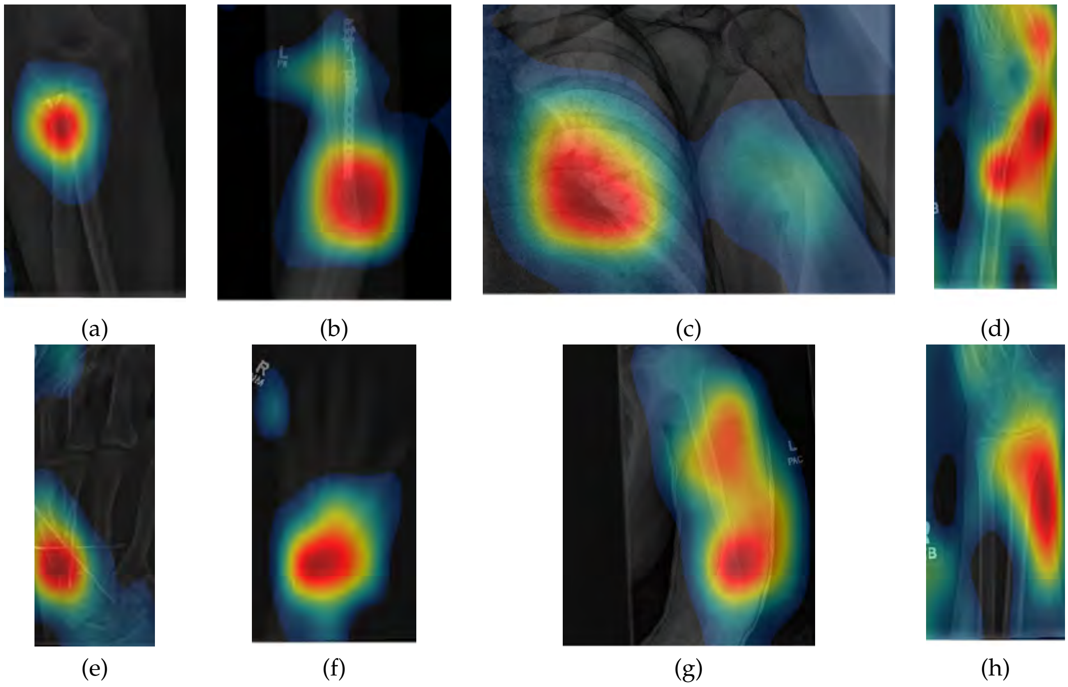

Figure 9. Illustration of Activation Maps overlaid over the eight radiographs with abnormalities of Figure 2 to indicate the regions of the image that activated a **ResNet 50** architecture. (a) Elbow, (b) Forearm, (c) Shoulder, (d) Wrist (lateral view), (d) Lateral view of Wrist, (e) Finger, (f) Hand, (g) Humerus, (h) Wrist. The activation maps illustrate the location of the abnormalities, e.g., (a,e), but appears spread in other cases (b,g) where the abnormality is detected together with a neighbouring region. In other cases (c) the abnormality is not detected.

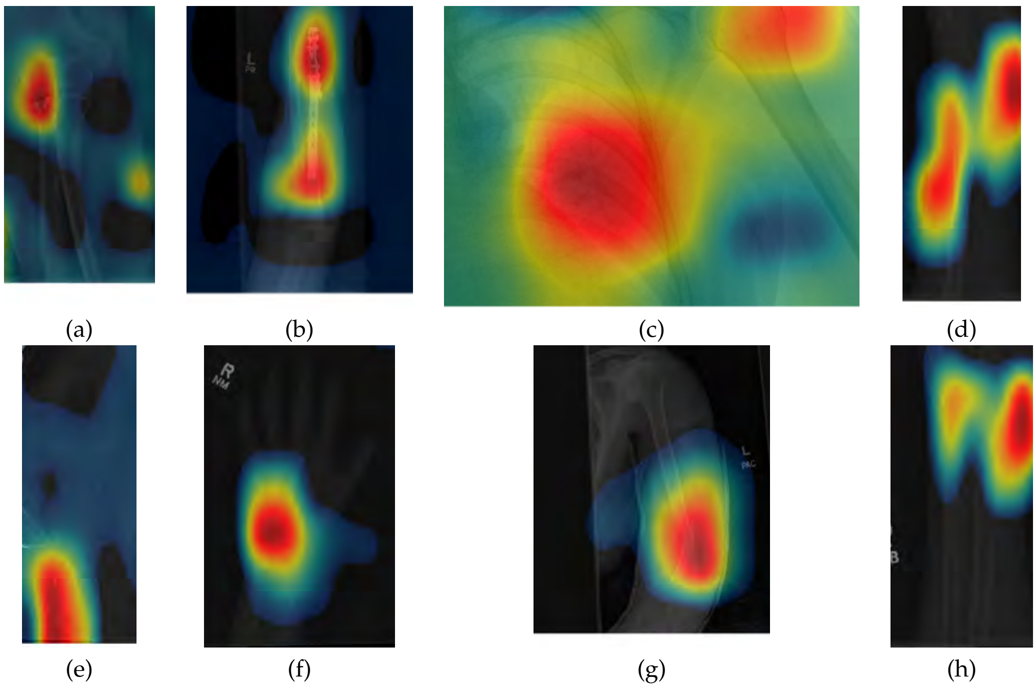

Figure 10. Illustration of Activation Maps overlaid over the eight radiographs with abnormalities of Figure 2 to indicate the regions of the image that activated an **Inception-Resnet-v2** architecture. As for the cases without abnormalities, the activation regions are more located e.g., (g) and in addition, the abnormalities are better located, e.g., (b,c).

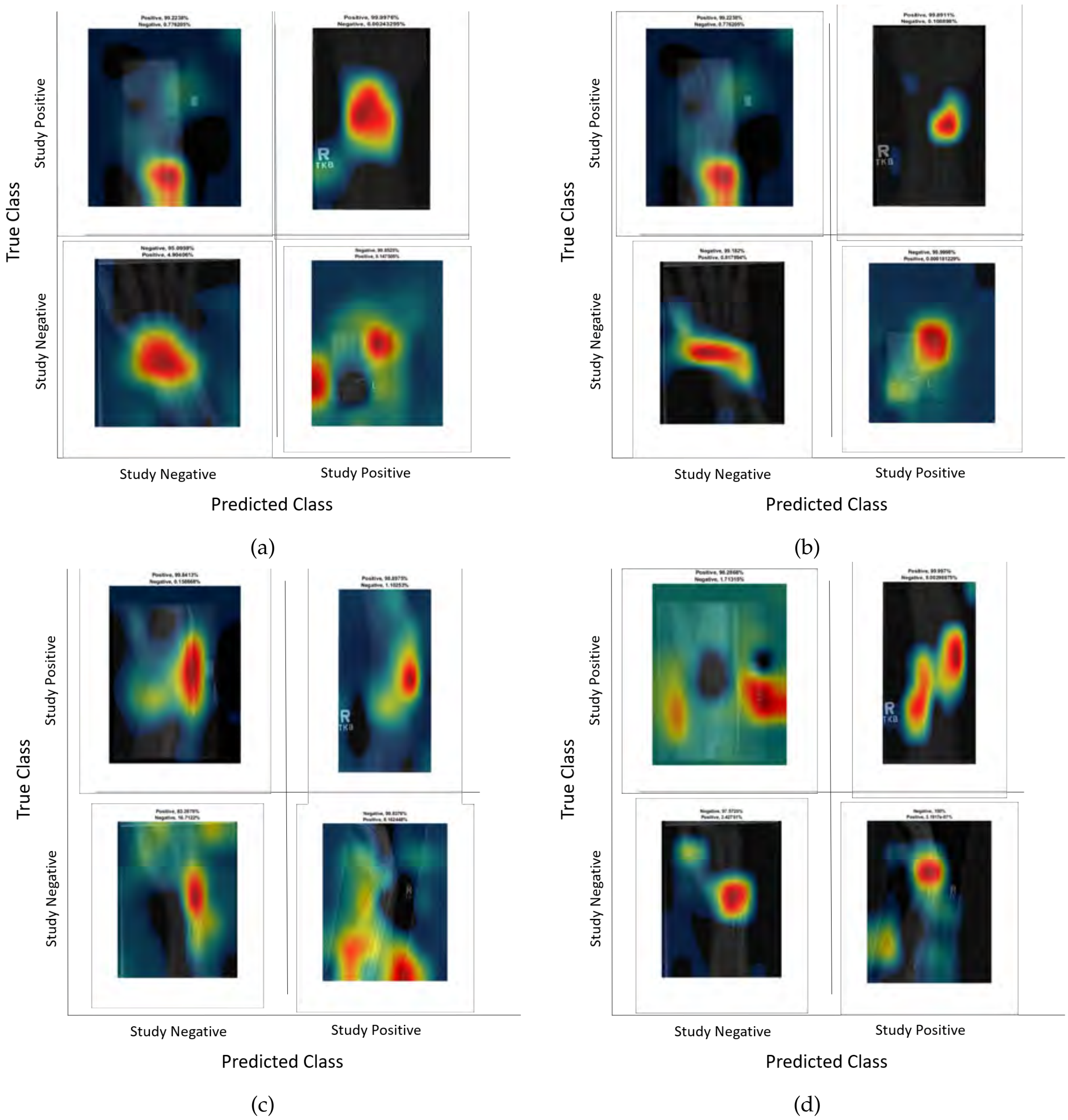

Figure 11. Illustration of the class Activation Maps overlaid on the four classification results for (a,b) Postero-anterior and (c,d) Lateral views shown in Figures[R1] 5,6 for ResNet 50 (a,c) and Inception-Resnet-v2 (b,d). In general Inception-Resnet-v2 presented more focused and smaller activation maps. It should also be noted that whilst for correct classifications, the highlighted regions are similar, for some incorrect classifications (c,d, top left and bottom right) the activations are quite different, which suggest that the architectures may not be confusing salient regions that are not related with the condition of normal or abnormal.[R1]

4. Discussion

In this paper, eleven CNN architectures for the classification of wrist x-rays were compared. Various hyper-parameters were attempted during the experiments. It was observed that Inception-Resnet-v2 provided the best results ($Ac = 0.747$, $\kappa = 0.548$), which were compared with leaders of the MURA challenge which reports 70 entries. The top three places of the leaderboard were $\kappa = 0.843, 0.834, 0.833$, the lowest score was $\kappa = 0.518$ and the best performance for a radiologist was $\kappa = 0.778$. Thus, without data augmentation, the results of all the networks were close to the bottom of the table. Data augmentation significantly improved the results to achieve the 25^{th} place of the

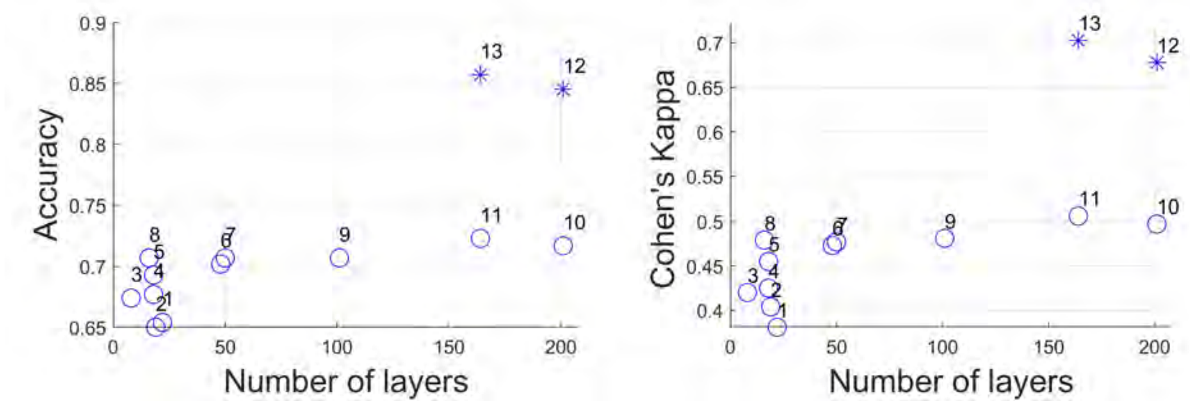

Figure 12. Illustration of the effect of the number of layers of architectures against the two metrics used in this paper Accuracy and Cohen's Kappa. Each architecture is represented by a circle, except those with augmentation that are represented by an asterisk. For visualisation purposes, numbers are added and these correspond to the order of Table 7 (1 GoogleNet, 2 VGG-19, 3 AlexNet, 4 SqueezeNet, 5 ResNet-18, 6 Inception-v3, 7 ResNet-50, 8 VGG-16, 9 ResNet-101, 10 DenseNet-201, 11 Inception-ResNet-v2, 12 ResNet-50 (augmentation), 13 Inception-ResNet-v2 (augmentation)). Notice the slight improvement provided by deeper networks and the significant improvement that corresponds to data augmentation.[R1]

leaderboard with ($Ac = 0.869$, $\kappa = 0.728$). Whilst this result was above the average of the table, the positive effect of data augmentation was confirmed to be close to human-level performance.

The CAM provides a channel to interpret how a CNN architecture is trained for feature extraction and the visualisation of the CAMs in the representative images was interesting in several aspects. First, the activated regions in ResNet-50 appeared more *broad-brushed* than those of the Inception-Resnet-v2. This applied both to the cases without abnormalities (Figures 7 and 8) and those with abnormalities (Figures 9 and 10); Second, the localisation of regions of attention by Inception-Resnet-v2 also appeared more precise than the ResNet-50. This can be appreciated in several cases, for instance the forearm that contains a metallic implant (b) and the humerus with a fracture (g); Third, the activation on the cases without abnormalities provides a consistent focus in areas where abnormalities are expected to appear. This suggests that the network has appropriately learned regions essential to the correct class prediction.

One important point to notice is that all the architectures provided lower results than those at the top of the MURA leaderboard table, even those tested with data augmentation. The top 3 architectures in the MURA leaderboard are: (1) base-comb2-xuan-v3 (ensemble) by *jzhang Availink*, (2) base-comb2-xuan (ensemble), also by *jzhang Availink* and (3) muti_type (ensemble model) by *SCU_ MILAB*. These reported the following Cohen's Kappa values of (1) 0.843, (2) 0.834 and (3) 0.833 respectively. Ensemble models are reported for the top 11 architectures and the highest single model is located in position 12 with a value of 0.773.[R3] Whilst in this paper only the wrist subset of the MURA dataset was analysed, it is not considered that these would be more difficult to classify than other anatomical parts. When data augmentation was applied to the input of the architectures, the results were significantly better, but still lower than the leaderboard winners. We speculate further steps could improve the performance of CNN-based classification. Specifically:

1. **Data Pre-Processing**: In addition to a grid search of the hyper-parameters, image pre-processing to remove irrelevant features (e.g. text labels) may help the network to target its attention. Appropriate data augmentations (e.g. rotation, reflection, etc) will allow better pattern recognition to be trained and, in turn, provides higher prediction accuracy.
2. **Post Training Evaluation**: Class Activation Map provides an interpretable visualisation for clinicians and radiologists to understand how a prediction was made. It allows the model to be re-trained with additional data to mitigate any model bias and discrepancy. Having a clear

association of the key features with the prediction classes [70] will aid in developing a more trustworthy CNN-based classification especially in a clinical setting.

3. **Model Ensemble** [71,72] or combination of the results of different architectures have also shown better results than an individual configuration. This is also observed in the leaderboard for the original MURA competition.
4. **Domain Knowledge**: The knowledge of anatomy (e.g. bone structure in elbow or hands [73]) or the location/orientation of bones [28] can be supplemented in a CNN-based classification to provide further fine tuning in anomaly detection as well as guiding the attention of the network for better results [74].

5. Conclusion

In this paper, an objective comparison of eleven convolutional neural networks was performed. The architectures were used to classify a large number of wrist radiographs which were divided into two groups, some that contained abnormalities, like fractures or metallic plates, and normal, i.e. healthy. The comparison showed a gradual improvement of the two metrics, namely, accuracy and Cohen's kappa, with more recent and deeper architectures. The best results were provided by ResNet-50 and Inception-Resnet-v2. Data augmentation was evaluated and was shown to increase the results significantly. Class activation maps were useful to observe the salient regions of each radiograph as they were passed through the architectures. Objective comparisons are important, especially for non-experts, who may consider one architecture without knowing if that is the optimal choice for their specific problem. [R3]

Supplementary Materials: The dataset for this study is publicly available by request from Stanford Machine Learning Group at https://stanfordmlgroup.github.io/competitions/mura/

Author Contributions: Conceptualization, A.A and C.C.R.-A.; Data curation, A.A; Formal analysis, A.A., K.H.N and C.C.R.-A.; Investigation, A.A., K.H.N and C.C.R.-A.; Methodology, A.A, K.H.N and C.C.R.-A.; Supervision, C.C.R.-A., E.A. and A.T.-S. Software, A.A., K.H.N and C.C.R.-A.; Writing—original draft preparation, A.A, K.H.N, C.K. and C.C.R.-A.; Writing—review and editing, A.A, K.H.N, C.K., A.T.S, E.A. and C.C.R.-A. All authors read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: A.A. has been supported through a Doctoral Scholarship by The School of Mathematics, Computer Science and Engineering at City, University of London. Dr. Karen Knapp from Exeter University is acknowledged her valuable discussions regarding fractures.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Abbreviations

The following abbreviations are used in this manuscript:

Ac	Accuracy
A&E	Accidents and Emergency
AI	Artificial Intelligence
CAM	Class Activation Mapping
CNN	Convolutional Neural Network
CT	Computed Tomography
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
MUA	Manipulation under Anaesthesia
MURA	Musculoskeletal Radiographs
ORIF	Open Reduction and Internal Fixation
ReLU	Rectified Linear Unit

References

1. Meena, S.; Sharma, P.; Sambharia, A.K.; Dawar, A. Fractures of Distal Radius: An Overview. *Journal of Family Medicine and Primary Care* **2014**, *3*, 325–332.
2. Raby, N.; Berman, L.; Morley, S.; De Lacey, G. *Accident and Emergency Radiology: A Survival Guide (Third Edition)*; Saunders Elsevier, 2015.
3. Bacorn, R.W.; Kurtzke, J.F. COLLES' FRACTURE: A Study of Two Thousand Cases from the New York State Workmen's Compensation Board. *JBJS* **1953**, *35*, 643–658.
4. Cooney, W.P.; Dobyns, J.H.; Linscheid, R.L. Complications of Colles' fractures. *The Journal of Bone and Joint Surgery. American Volume* **1980**, *62*, 613–619.
5. Vergara, I.; Vrotsou, K.; Orive, M.; Garcia-Gutierrez, S.; Gonzalez, N.; Las Hayas, C.; Quintana, J.M. Wrist fractures and their impact in daily living functionality on elderly people: a prospective cohort study. *BMC geriatrics* **2016**, *16*, 11.
6. Diaz-Garcia, R.J.; Chung, K.C. The Evolution of Distal Radius Fracture Management – A Historical Treatise. *Hand Clinics* **2012**, *28*, 105–111.
7. Redfern, R. A regional examination of surgery and fracture treatment in Iron Age and Roman Britain. *International Journal of Osteoarchaeology* **2010**, *20*, 443–471. doi:https://doi.org/10.1002/oa.1067.
8. Barai, A.; Lambie, B.; Cosgrave, C.; Baxter, J. Management of distal radius fractures in the emergency department: A long-term functional outcome measure study with the Disabilities of Arm, Shoulder and Hand (DASH) scores. *Emergency Medicine Australasia* **2018**, *30*, 530–537, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/1742-6723.12946]. doi:10.1111/1742-6723.12946.
9. Malik, H.; Appelboam, A.; Taylor, G. Colles' type distal radial fractures undergoing manipulation in the ED: a multicentre observational cohort study. *Emergency medicine journal: EMJ* **2020**, *37*, 498–501. doi:10.1136/emermed-2020-209478.
10. Arora, R.; Gabl, M.; Gschwentner, M.; Deml, C.; Krappinger, D.; Lutz, M. A Comparative Study of Clinical and Radiologic Outcomes of Unstable Colles Type Distal Radius Fractures in Patients Older Than 70 Years: Nonoperative Treatment Versus Volar Locking Plating. *Journal of Orthopaedic Trauma* **2009**, *23*, 237–242. doi:10.1097/BOT.0b013e31819b24e9.
11. Sellbrandt, I.; Brattwall, M.; Warrén Stomberg, M.; Jildenstål, P.; Jakobsson, J.G. Anaesthesia for open wrist fracture surgery in adults/elderly. *F1000Research* **2017**, *6*, 1996. doi:10.12688/f1000research.13004.1.
12. Dukan, R.; Krief, E.; Nizard, R. Distal radius fracture volar locking plate osteosynthesis using wide-awake local anaesthesia. *The Journal of Hand Surgery, European Volume* **2020**, *45*, 857–863. doi:10.1177/1753193420916418.
13. Arora, R.; Lutz, M.; Hennerbichler, A.; Krappinger, D.; Md, D.E.; Gabl, M. Complications Following Internal Fixation of Unstable Distal Radius Fracture With a Palmar Locking-Plate. *Journal of Orthopaedic Trauma* **2007**, *21*, 316–322. doi:10.1097/BOT.0b013e318059b993.
14. Gaspar, M.P.; Lou, J.; Kane, P.M.; Jacoby, S.M.; Osterman, A.L.; Culp, R.W. Complications Following Partial and Total Wrist Arthroplasty: A Single-Center Retrospective Review. *Journal of Hand Surgery* **2016**, *41*, 47–53.e4.
15. Bartl, C.; Stengel, D.; Bruckner, T.; Rossion, I.; Luntz, S.; Seiler, C.; Gebhard, F. Open reduction and internal fixation versus casting for highly comminuted and intra-articular fractures of the distal radius (ORCHID): protocol for a randomized clinical multi-center trial. *Trials* **2011**, *12*, 84. doi:10.1186/1745-6215-12-84.
16. Grewal, R.; MacDermid, J.C.; King, G.J.W.; Faber, K.J. Open Reduction Internal Fixation Versus Percutaneous Pinning With External Fixation of Distal Radius Fractures: A Prospective, Randomized Clinical Trial. *Journal of Hand Surgery* **2011**, *36*, 1899–1906. doi:10.1016/j.jhsa.2011.09.015.
17. Kapoor, H.; Agarwal, A.; Dhaon, B.K. Displaced intra-articular fractures of distal radius: a comparative evaluation of results following closed reduction, external fixation and open reduction with internal fixation. *Injury* **2000**, *31*, 75–79. doi:10.1016/S0020-1383(99)00207-7.
18. Kelly, A.J.; Warwick, D.; Crichlow, T.P.K.; Bannister, G.C. Is manipulation of moderately displaced Colles' fracture worthwhile? A prospective randomized trial. *Injury* **1997**, *28*, 283–287.
19. Handoll, H.H.; Madhok, R. Conservative interventions for treating distal radial fractures in adults. *Cochrane Database of Systematic Reviews* **2003**. doi:10.1002/14651858.CD000314.

20. Handoll, H.H.; Madhok, R. Closed reduction methods for treating distal radial fractures in adults. *Cochrane Database of Systematic Reviews* **2003**. doi:10.1002/14651858.CD003763.
21. Handoll, H.H.; Huntley, J.S.; Madhok, R. Different methods of external fixation for treating distal radial fractures in adults. *Cochrane Database of Systematic Reviews* **2008**. doi:10.1002/14651858.CD006522.pub2.
22. NHS Statistics. Statistics: Diagnostic Imaging Dataset 2018-19 Data, https://www.england.nhs.uk /statistics/statistical-work-areas/diagnostic-imaging-dataset/ diagnostic-imaging-dataset-2018-19-data/.
23. The Royal College of Radiologists. The NHS does not have enough radiologists to keep patients safe, say three-in-four hospital imaging bosses, https://www.rcr.ac.uk/posts/ nhs-does-not-have-enough-radiologists-keep-patients-safe-say-three-four-hospital-imaging.
24. Lee, J.I.; Park, K.C.; Joo, I.H.; Jeong, H.W.; Park, J.W. The Effect of Osteoporosis on the Outcomes After Volar Locking Plate Fixation in Female Patients Older than 50 Years With Unstable Distal Radius Fractures. *The Journal of Hand Surgery* **2018**, *43*, 731–737. doi:10.1016/j.jhsa.2018.05.028.
25. Wang, J.; Lu, Y.; Cui, Y.; Wei, X.; Sun, J. Is volar locking plate superior to external fixation for distal radius fractures? A comprehensive meta-analysis. *Acta Orthopaedica Et Traumatologica Turcica* **2018**, *52*, 334–342. doi:10.1016/j.aott.2018.06.001.
26. Sharareh, B.; Mitchell, S. Radiographic Outcomes of Dorsal Spanning Plate for Treatment of Comminuted Distal Radius Fractures in Non-Elderly Patients. *Journal of Hand Surgery Global Online* **2020**, *2*, 94–101. doi:10.1016/j.jhsg.2019.10.001.
27. Rhee, S.H.; Kim, J. Distal radius fracture metaphyseal comminution: A new radiographic parameter for quantifying, the metaphyseal collapse ratio (MCR). *Orthopaedics & Traumatology: Surgery & Research* **2013**, *99*, 713–718. doi:10.1016/j.otsr.2013.05.002.
28. Reyes-Aldasoro, C.C.; Ngan, K.H.; Ananda, A.; Garcez, A.d.; Appelboam, A.; Knapp, K.M. Geometric semi-automatic analysis of radiographs of Colles' fractures. *PLOS ONE* **2020**, *15*, e0238926. doi:10.1371/journal.pone.0238926.
29. Adolphson, P.; Abbaszadegan, H.; Jonsson, U. Computer-assisted prediction of the instability of Colles' fractures. *International Orthopaedics* **1993**, *17*, 13–15. doi:10.1007/bf00195215.
30. Erhart, S.; Toth, S.; Kaiser, P.; Kastenberger, T.; Deml, C.; Arora, R. Comparison of volarly and dorsally displaced distal radius fracture treated by volar locking plate fixation. *Archives of Orthopaedic and Trauma Surgery* **2018**, *138*, 879–885. doi:10.1007/s00402-018-2925-x.
31. Zenke, Y.; Furukawa, K.; Furukawa, H.; Maekawa, K.; Tajima, T.; Yamanaka, Y.; Hirasawa, H.; Menuki, K.; Sakai, A. Radiographic Measurements as a Predictor of Correction Loss in Conservative Treatment of Colles' Fracture. *Journal of UOEH* **2019**, *41*, 139–144. doi:10.7888/juoeh.41.139.
32. Rabar, S.; Lau, R.; O'Flynn, N.; Li, L.; Barry, P. Risk assessment of fragility fractures: summary of NICE guidance. *BMJ* **2012**, *345*. doi:10.1136/bmj.e3698.
33. Knapp, K.M.; Meertens, R.M.; Seymour, R. Imaging and opportunistic identification of fractures. *Pavilion Publishing* **2018**, *Vol.48(11)*, 10–12.
34. Crespo, R.; Revilla, M.; Usabiago, J.; Crespo, E.; García-Ariño, J.; Villa, L.F.; Rico, H. Metacarpal Radiogrammetry by Computed Radiography in Postmenopausal Women with Colles' Fracture and Vertebral Crush Fracture Syndrome. *Calcified Tissue International* **1998**, *62*, 470–473. doi:10.1007/s002239900463.
35. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016. http://www.deeplearningbook.org.
36. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.
37. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2013**, *35*, 1798–1828. doi:10.1109/TPAMI.2013.50.
38. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A.C.; Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* **2015**, *115*, 211–252. doi:10.1007/s11263-015-0816-y.
39. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. doi:10.1038/nature21056.

40. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *39*, 2481–2495. doi:10.1109/TPAMI.2016.2644615.
41. Chen, C.; Qin, C.; Qiu, H.; Tarroni, G.; Duan, J.; Bai, W.; Rueckert, D. Deep learning for cardiac image segmentation: A review. *arXiv e-prints* **2019**, p. arXiv:1911.03723, [arXiv:eess.IV/1911.03723].
42. Meyer, P.; Noblet, V.; Mazzara, C.; Lallement, A. Survey on deep learning for radiotherapy. *Computers in Biology and Medicine* **2018**, *98*, 126–146.
43. Gibson, E.; Li, W.; Sudre, C.; Fidon, L.; Shakir, D.I.; Wang, G.; Eaton-Rosen, Z.; Gray, R.; Doel, T.; Hu, Y.; et al.. NiftyNet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine* **2018**, *158*, 113–122.
44. Iglesias, J.; Sabuncu, M. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis* **2015**, *24*, 205–219.
45. Siuly, S.; Zhang, Y. Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis. *Data Science and Engineering* **2016**, *1*, 54–64.
46. Kather, J.N.; Krisam, J.; Charoentong, P.; Luedde, T.; Herpel, E.; Weis, C.A.; Gaiser, T.; Marx, A.; Valous, N.A.; Ferber, D.; et al.. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine* **2019**, *16*, e1002730. doi:10.1371/journal.pmed.1002730.
47. Luo, J.; Wu, M.; Gopukumar, D.; Zhao, Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights* **2016**, *8*, BII.S31559.
48. Viceconti, M.; Hunter, P.; Hose, R. Big Data, Big Knowledge: Big Data for Personalized Healthcare. *IEEE Journal of Biomedical and Health Informatics* **2015**, *19*, 1209–1215. doi:10.1109/JBHI.2015.2406883.
49. Blüthgen, C.; Becker, A.S.; Vittoria de Martini, I.; Meier, A.; Martini, K.; Frauenfelder, T. Detection and localization of distal radius fractures: Deep learning system versus radiologists. *European Journal of Radiology* **2020**, *126*, 108925. doi:10.1016/j.ejrad.2020.108925.
50. Lindsey, R.; Daluiski, A.; Chopra, S.; Lachapelle, A.; Mozer, M.; Sicular, S.; Hanel, D.; Gardner, M.; Gupta, A.; Hotchkiss, R.; et al.. Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences* **2018**, *115*, 11591–11596. doi:10.1073/pnas.1806905115.
51. Thian, Y.L.; Li, Y.; Jagmohan, P.; Sia, D.; Chan, V.E.Y.; Tan, R.T. Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs. *Radiology: Artificial Intelligence* **2019**, *1*, e180001. doi:10.1148/ryai.2019180001.
52. Castelvecchi, D. Can we open the black box of AI? *Nature News* **2016**, *538*, 20. doi:10.1038/538020a.
53. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **2019**, *1*, 206–215. doi:10.1038/s42256-019-0048-x.
54. Zednik, C. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *arXiv:1903.04361 [cs]* **2019**. arXiv: 1903.04361.
55. Aggarwal, A.; Lohia, P.; Nagar, S.; Dey, K.; Saha, D. Black box fairness testing of machine learning models. Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. Association for Computing Machinery, 2019, ESEC/FSE 2019, p. 625–635. doi:10.1145/3338906.3338937.
56. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p. 2921–2929. doi:10.1109/CVPR.2016.319.
57. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* **2020**, *128*, 336–359. doi:10.1007/s11263-019-01228-7.
58. Ananda.; Karabag, C.; Ter-Sarkisov, A.; Alonso, E.; Reyes-Aldasoro, C.C. Radiography Classification: A Comparison between Eleven Convolutional Neural Networks. 2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA), 2020, p. 119–125. doi:10.1109/MCNA50957.2020.9264285.
59. Rajpurkar, P.; Irvin, J.; Bagul, A.; Ding, D.; Duan, T.; Mehta, H.; Yang, B.; Zhu, K.; Laird, D.; Ball, R.L.; et al.. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. *arXiv:1712.06957* **2017**.

60. Krizhevsky, A.; Sutskever, I.; Hinton, G.E., ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*; Pereira, F.; Burges, C.J.C.; Bottou, L.; Weinberger, K.Q., Eds.; Curran Associates, Inc., 2012; p. 1097–1105.
61. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, p. 1–9. doi:10.1109/CVPR.2015.7298594.
62. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* **2015**. arXiv: 1409.1556.
63. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
64. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567 [cs]* **2015**. arXiv: 1512.00567.
65. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv:1602.07360 [cs]* **2016**. arXiv: 1602.07360.
66. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs]* **2016**. arXiv: 1608.06993.
67. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv:1602.07261 [cs]* **2016**. arXiv: 1602.07261.
68. Pizer, S.M.; Johnston, R.E.; Ericksen, J.P.; Yankaskas, B.C.; Muller, K.E. Contrast-limited adaptive histogram equalization: speed and effectiveness. IEEE Computer Society, 1990, pp. 337,338,339,340,341,342,343,344,345–337,338,339,340,341,342,343,344,345. doi:10.1109/VBC.1990.109340.
69. McHugh, M.L. Interrater reliability: the kappa statistic. *Biochemia Medica* **2012**, *22*, 276–282.
70. Oramas, J.; Wang, K.; Tuytelaars, T. Visual Explanation by Interpretation: Improving Visual Feedback Capabilities of Deep Neural Networks. International Conference on Learning Representations, 2019.
71. Derkatch, S.; Kirby, C.; Kimelman, D.; Jozani, M.J.; Davidson, J.M.; Leslie, W.D. Identification of Vertebral Fractures by Convolutional Neural Networks to Predict Nonvertebral and Hip Fractures: A Registry-based Cohort Study of Dual X-ray Absorptiometry. *Radiology* **2019**, *293*, 405–411. doi:10.1148/radiol.2019190201.
72. Mondol, T.C.; Iqbal, H.; Hashem, M. Deep CNN-Based Ensemble CADx Model for Musculoskeletal Abnormality Detection from Radiographs. 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), 2019, p. 392–397. doi:10.1109/ICAEE48663.2019.8975455.
73. Chen, X.; Graham, J.; Hutchinson, C.; Muir, L. Automatic Inference and Measurement of 3D Carpal Bone Kinematics From Single View Fluoroscopic Sequences. *IEEE Transactions on Medical Imaging* **2013**, *32*, 317–328. doi:10.1109/TMI.2012.2226740.
74. Xie, X.; Niu, J.; Liu, X.; Chen, Z.; Tang, S.; Yu, S. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis* **2021**, *69*, 101985. doi:10.1016/j.media.2021.101985.